普通高等教育“十三五”规划教材

◎ 何晓群　编著

应用回归分析

Applied Regression Analysis

（R Language Edition）

（R语言版）

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

普通高等教育“十三五”规划教材

◎ 何晓群　编著

应用回归分析

Applied Regression Analysis

(R Language Edition)

（R语言版）

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

回归分析是统计学中一个非常重要的分支，在自然科学、管理及社会经济等领域有着非常广泛的应用。本书是针对统计学专业和财经管理类专业教学的需要而编写的。

本书写作的指导思想是在不失严谨的前提下，明显不同于纯数理类教材，努力突出实际案例的应用和统计思想的渗透。由于 R 语言已风靡全球，在统计方法的应用中运用 R 语言也被越来越多的中国学者所追捧，因此本书结合 R 软件全面系统地介绍回归分析的实用方法，尽量结合中国社会经济、自然科学等领域的研究实例，把回归分析的方法与实际应用结合起来，注重定性分析与定量分析的紧密结合，努力把同行以及我们在实践中应用回归分析的经验和体会融入其中。

本书既可作为统计学、应用统计学和经济统计学三个本科专业的回归分析课程教材，还可作为非统计专业研究生现代统计分析方法与应用及定量分析与建模课程的教材，同时也适合有意学习 R 语言和回归建模技术的实际工作者阅读和参考。

图书在版编目 (CIP) 数据

应用回归分析：R 语言版 / 何晓群编著. — 北京：电子工业出版社，2017.7

ISBN 978-7-121-31652-4

I. ①应… II. ①何… III. ①回归分析—高等学校—教材 IV. ①O212.1

中国版本图书馆 CIP 数据核字(2017)第 122203 号

策划编辑：王志宇
责任编辑：王志宇
印　　刷：北京中新伟业印刷有限公司
装　　订：北京中新伟业印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787 × 1 092　1/16　印张：17.75　字数：400 千字　插页：1
版　　次：2017 年 7 月第 1 版
印　　次：2017 年 7 月第 1 次印刷
定　　价：42.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010) 88254523，wangzy@phei.com.cn。

前　言

回归分析是统计学中一个非常重要的分支，在自然科学、管理科学和社会经济等领域有着非常广泛的应用。本书是针对统计学专业和财经管理类专业教学的需要而编写的。

本书写作的指导思想是在不失严谨的前提下，明显不同于纯数理类教材，努力突出实际案例的应用和统计思想的渗透，结合R软件全面系统地介绍回归分析的使用方法，尽量结合中国社会经济、自然科学等领域的研究实例，把回归分析的方法与实际应用结合起来，注重定性分析与定量分析的紧密结合，努力把同行以及我们在实践中应用回归分析的经验和体会融入其中。

全书分为10章。第1章对回归分析的研究内容和建模过程给出综述性介绍；第2、3章详细介绍了一元和多元线性回归的参数估计、显著性检验及其应用；第4章对违背回归模型基本假设的异方差、自相关和异常值等问题给出了诊断和处理方法；第5章介绍了回归变量选择与逐步回归方法；第6章就多重共线性的产生背景、诊断方法、处理方法等方面结合实际经济问题进行了讨论；第7章岭回归估计是解决共线性问题的一种非常实用的方法；第8章介绍了主成分回归与偏最小二乘；第9章介绍了可化为线性回归的曲线回归、多项式回归，以及不能线性化的非线性回归模型的计算；第10章分别介绍了自变量中含定性变量和因变量是定性变量的回归问题，以及因变量是多类别和有序变量的回归问题。

本书作为回归分析的应用性教材，讲述的重点是结合R语言软件实现回归分析中的各种方法，比较各种方法的适用条件，并解释分析结果。为了保持教材的完整性，对一些基本的公式和定理给出了推导和证明，对有些基本的理论及性质也做了必要的说明。书后习题用R语言来完成，为了节省篇幅本书只给出习题答案的简要内容，详细答案内容及有关R代码我们放在中国人民大学六西格玛质量管理研究中心网站供需求者下载(www.ruc-6sigma.com)。

对于统计学专业的本科生可以全面系统地讲述本教材的内容；对非统计学专业的本科生应该舍弃其中理论性质的内容；对非统计学专业的研究生可以根据具体情况选择讲授其中的内容。根据我们的教学实践，本书讲授51课时较为合适，若有多媒体设备的配合，教学将会更为方便和有效。

我的博士研究生刘赛可、王蕾、夏利宇为本书编写做了全面的上机实践。本书的大部分例题是我们多年教学和科研工作的积累，部分实例为体现其典型性引用了他人著作。在此谨向对本书出版提供帮助的师长和朋友表示衷心的感谢。

由于水平所限，书中难免有不足之处，尤其是在一些应用研究的体会性讨论中，恐有偏颇之处，恳切希望读者批评指正。

何晓群

于中国人民大学统计学院

中国人民大学应用统计科学研究中心

目　录

第 1 章

回归分析概述

为了在系统学习回归分析之前对该课程的思想方法、主要内容、发展现状等有个概括的了解，本章将由变量间的统计关系引申出社会经济与自然科学等现象中的相关与回归问题，并扼要介绍“回归”名称的由来及近代回归分析的发展、回归分析研究的主要内容，以及建立回归模型的步骤与建模过程中应注意的问题。

1.1 变量间的相关关系

社会经济与自然科学等现象之间的相互联系和制约是一个普遍规律。例如社会经济的发展总是与一定的经济变量的数量变化紧密联系着。社会经济现象不仅同和它有关的现象构成一个普遍联系的整体，而且在它的内部存在着许多彼此关联的因素，在一定的社会环境、地理条件、政府决策影响下，一些因素推动或制约另外一些与之联系的因素发生变化。这种状况表明，在经济现象的内部和外部联系中存在着一定的相关性，人们往往利用这种相关关系来制定有关的经济政策，以指导、控制社会经济活动的发展。要认识和掌握客观经济规律就必须探求经济现象中经济变量的变化规律，变量间的统计关系是经济变量变化规律的重要特征。

互有联系的经济现象及经济变量间关系的紧密程度各不一样。一种极端的情况是一个变量的变化能完全决定另一个变量的变化。例如，一家保险公司承保汽车 5 万辆，每辆保费收入为 1 000 元，则该保险公司汽车承保总收入为 5 000 万元。如果把承保总收入记为 y，承保汽车辆数记为 x，则 $y=1\,000x$。x 与 y 两个变量间完全表现为一种确定性关系，即函数关系，如图 1-1 所示。

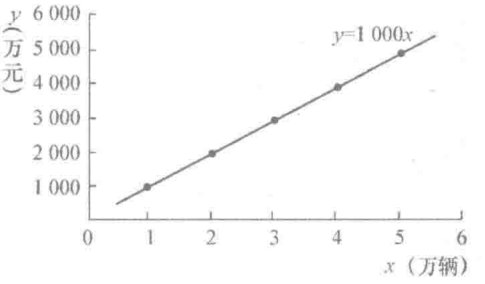

图 1-1　函数关系图

又如，银行的一年期存款利率为 2.55%，存入的本金用 x 表示，到期的本息用 y 表示，则 $y=x+2.55\%x$。这里 y 与 x 仍表现为一种函数关系。对于任意两个变量间的函数关系，可以表述为下面的数学形式

$$y = f(x)$$

再如，工业企业的原材料消耗总额用 y 表示，生产量用 x_1 表示，单位产量消耗用 x_2 表示，原材料价格用 x_3 表示，则

$$y = x_1 x_2 x_3$$

这里的 y 与 x_1，x_2，x_3 仍是一种确定性的函数关系，但它们显然不是线性函数关系。我们可以将变量 y 与 p 个变量 $x_1, x_2, \cdots, x_p$ 之间存在的某种函数关系用下面的形式表示

$$y = f(x_1, x_2, \cdots, x_p)$$

经济问题中还有很多函数关系的例子。物理学中的自由落体距离公式、初等数学中的许多计算公式等表示的都是变量间的函数关系。

然而，现实世界中还有不少情况是两事物之间有着密切的联系，但它们密切的程度并没有到由一个可以完全确定另一个的地步，下面举几个例子。

(1)我们都知道某种高档消费品的销售量与城镇居民的收入密切相关，居民收入高，这种消费品的销售量就大。但是由居民收入 x 并不能完全确定某种高档消费品的销售量 y，因为这种高档消费品的销售量还受人们的消费习惯、心理因素、其他商品的吸引程度及价格的高低等诸多因素的影响。这样变量 y 与变量 x 就是一种非确定的关系，如图 1-2 所示。

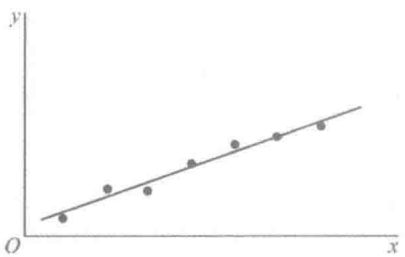

图 1-2　y 与 x 非确定性关系图

(2)粮食产量 y 与施肥量 x 之间有密切的关系，在一定的范围内，施肥量越多，粮食产量就越高。但是，施肥量并不能完全确定粮食产量，因为粮食产量还与其他因素有关，如降雨量、田间管理水平等。因此粮食产量 y 与施肥量 x 之间不存在确定的函数关系。

(3)储蓄额与居民的收入密切相关，但是由居民收入并不能完全确定储蓄额。因为影响储蓄额的因素很多，如通货膨胀、股票价格指数、利率、消费观念、投资意识等。因此尽管储蓄额与居民收入有密切的关系，但它们之间并不存在一种确定性关系。

再如广告费支出与商品销售额、保险利润与保费收入、工业产值与用电量等。这方面的例子不胜枚举。

以上变量间关系的一个共同特征是尽管密切，但却是一种非确定性关系。由于经济问题的复杂性，有许多因素因为我们的认识以及其他客观原因的局限，并没有包含在内，或者由于试验误差、测量误差以及其他种种偶然因素的影响，使得另外一个或一些变量的取值带有一定的随机性。因此当一个或一些变量取定值后，不能以确定值与之对应。

从图 1-1 看到确定性的函数关系，各对应点完全落在一条直线上。而由图 1-2 看到，各对应点并不完全落在一条直线上，即有的点在直线上，有的点在直线的两侧。这种对应点不能分布在一条直线上的变量间的关系，也就是变量 x 与 y 之间有一定的关系，但是又没有密切到可以通过 x 唯一确定 y 的程度，这种关系正是统计学研究的

重要内容。在推断统计中，我们把上述变量间具有密切关联而又不能由某一个或某一些变量唯一确定另外一个变量的关系称为变量间的统计关系或相关关系。这种统计关系的规律性是统计学中研究的主要对象，现代统计学中关于统计关系的研究已形成两个重要的分支，它们叫回归分析和相关分析。

回归分析和相关分析都是研究变量间关系的统计学课题。在应用中，两种分析方法经常相互结合和渗透，但它们研究的侧重点和应用面不同。它们的差别主要有以下几点：一是在回归分析中，变量 y 称为因变量，处在被解释的特殊地位。在相关分析中，变量 y 与变量 x 处于平等的地位，即研究变量 y 与变量 x 的密切程度与研究变量 x 与变量 y 的密切程度是一回事。二是相关分析中所涉及的变量 y 与 x 全是随机变量。而回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量。通常的回归模型中，我们总是假定 x 是非随机的确定变量。三是相关分析的研究主要是为刻画两类变量间线性相关的密切程度。而回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制。

由于回归分析与相关分析研究的侧重点不同，它们的研究方法也大不相同。回归分析已成为现代统计学中应用最广泛、研究最活跃的一个独立分支。

1.2　“回归”思想及名称的由来

回归分析是处理变量 x 与 y 之间的关系的一种统计方法和技术。这里所研究的变量之间的关系就是上述的统计关系，即当给定 x 的值，y 的值不能确定，只能通过一定的概率分布来描述。于是，我们称给定 x 时 y 的条件数学期望

$$f(x)=E(y|x) \tag{1.1}$$

为随机变量 y 对 x 的回归函数，或称为随机变量 y 对 x 的均值回归函数。式(1.1)从平均意义上刻画了变量 x 与 y 之间的统计规律。

在实际问题中，我们把 x 称为自变量，y 称为因变量。如果要由 x 预测 y，就是要利用 x，y 的观察值，即样本观测值

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n) \tag{1.2}$$

来建立一个函数，当给定 x 值后，代入此函数中算出一个 y 值，这个值就称为 y 的预测值。如何建立这个函数？这就要从样本观测值 (x_i, y_i) 出发，观察 (x_i, y_i) 在平面直角坐标系上的分布情况，图 1-2 就是居民收入与商品销售量的散点图。由这个图可看出样本点基本上分布在一条直线的周围，因而要确定商品销售量 y 与居民收入 x 的关系，可考虑用一个线性函数来描述。图 1-2 中的直线即线性方程

$$E(y|x)=\alpha+\beta x \tag{1.3}$$

方程式(1.3)中的参数 α，β 尚不知道，这就需要由样本数据(1.2)去进行估计。具

体如何估计参数 α, β，我们将在第 2 章中详细介绍。

当我们由样本数据(1.2)估计出 α, β 的值后，用估计值 $\hat{\alpha}, \hat{\beta}$ 分别代替式(1.3)中的 α, β，得方程

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{1.4}$$

方程式(1.4)就称为回归方程。这里因为因变量 y 与自变量 x 呈线性关系，故称式(1.4)为 y 对 x 的线性回归方程。又因式(1.4)的建立依赖于观察或试验积累的数据(1.2)，所以又称式(1.4)为经验回归方程。相对这种叫法，我们把式(1.3)称为理论回归方程。理论回归方程是设想把所研究问题的总体中每一个体的 (x, y) 值都测量了，利用其全部测量结果而建立的回归方程，这在实际中是做不到的。理论回归方程中的 α 是方程式(1.3)所画出的直线在 y 轴上的截距，β 为直线的斜率，它们分别称为回归常数和回归系数。而方程式(1.4)中的参数 $\hat{\alpha}, \hat{\beta}$ 称为经验回归常数和经验回归系数。

回归分析的基本思想和方法以及“回归”(regression)名称的由来归功于英国统计学家 F.高尔顿(F.Galton，1822—1911)。高尔顿和他的学生、现代统计学的奠基者之一 K.皮尔逊(K.Pearson，1856—1936)在研究父母身高与其子女身高的遗传问题时，观察了 1078 对夫妇，以每对夫妇的平均身高作为 x，而取他们的一个成年儿子的身高作为 y，将结果在平面直角坐标系上绘成散点图，发现趋势近乎一条直线。计算出的回归直线方程为

$$\hat{y} = 33.73 + 0.516x \tag{1.5}$$

这种趋势及回归方程总的表明父母平均身高 x 每增加一个单位，其成年儿子的身高 y 平均增加 0.516 个单位。这个结果表明，虽然高个子父辈的确有生高个子儿子的趋势，但父辈身高增加一个单位，儿子身高仅增加半个单位左右。反之，矮个子父辈的确有生矮个子儿子的趋势，但父辈身高减少一个单位，儿子身高仅减少半个单位左右。通俗地说，一群特高个子父辈(例如排球运动员)的儿子们在同龄人中平均仅为高个子，一群高个子父辈的儿子们在同龄人中平均仅为略高个子；一群特矮个子父辈的儿子们在同龄人中平均仅为矮个子，一群矮个子父辈的儿子们在同龄人中平均仅为略矮个子，即子代的平均高度向中心回归了。正是因为子代的身高有回到同龄人平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象。这个例子生动地说明了生物学中“种”的概念的稳定性。正是为了描述这种有趣的现象，高尔顿引进了“回归”这个名词来描述父辈身高 x 与子辈身高 y 的关系。尽管“回归”这个名称的由来具有其特定的含义，而在人们研究的大量问题中，其变量 x 与 y 之间的关系并不总是具有这种“回归”的含义，但仍借用这个名词把研究变量 x 与 y 间统计关系的量化方法称为“回归”分析，也算是对高尔顿这位伟大的统计学家的纪念。

1.3　回归分析的主要内容及其一般模型

1.3.1　回归分析研究的主要内容

回归分析研究的主要对象是客观事物变量间的统计关系，它是建立在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在那些看上去是不确定的现象中的统计规律性的统计方法。回归分析方法是通过建立统计模型研究变量间相互关系的密切程度、结构状态及进行模型预测的一种有效的工具。

回归分析方法在生产实践中的广泛应用是其发展和完善的根本动力。如果从 19 世纪初(1809 年)高斯(Gauss)提出最小二乘法算起，回归分析的历史已有 200 多年。从经典的回归分析方法到近代的回归分析方法，它们所研究的内容已非常丰富。如果按研究的方法来划分，回归分析研究的范围大致如下：

- 回归分析
  - 线性回归
    - 一元线性回归
    - 多元线性回归
    - 多个因变量与多个自变量的回归
  - 回归诊断
    - 讨论如何从数据推断回归模型基本假设的合理性
    - 当基本假设不成立时如何对模型进行修正
    - 判定回归方程拟合的效果
    - 选择回归函数的形式
  - 回归变量的选择
    - 自变量选择的准则
    - 逐步回归分析方法
  - 参数估计方法的改进
    - 岭回归
    - 主成分回归
    - 偏最小二乘法
  - 非线性回归
    - 一元非线性回归
    - 分段回归
    - 多元非线性回归
  - 含有定性变量的回归
    - 自变量含定性变量的情况
    - 因变量是定性变量的情况

1.3.2　回归模型的一般形式

如果变量 x_1, x_2,…, x_p 与随机变量 y 之间存在着相关关系，通常就意味着每当 x_1, x_2,…, x_p 取值确定后，y 便有相应的概率分布与之对应。随机变量 y 与相关变量 x_1, x_2,…, x_p 之间的模型为

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon \tag{1.6}$$

式中，随机变量 y 称为被解释变量(因变量)；$x_1, x_2, \cdots, x_p$ 称为解释变量(自变量)。在计量经济学中，也称因变量为内生变量，自变量为外生变量。$f(x_1, x_2, \cdots, x_p)$ 为一般变量 $x_1, x_2, \cdots, x_p$ 的确定性关系；ε 为随机误差。正是因为随机误差项 ε 的引入，才将变量之间的关系描述为一个随机方程，使得我们可以借助随机数学方法研究 y 与 $x_1, x_2, \cdots, x_p$ 的关系。由于客观经济现象是错综复杂的，一种经济现象很难用有限个因素来准确说明，随机误差项可以概括表示由于人们的认识以及其他客观原因的局限而没有考虑的种种偶然因素。随机误差项主要包括下列因素的影响：

(1)由于人们认识的局限或时间、费用、数据质量等的制约未引入回归模型但又对回归被解释变量 y 有影响的因素。

(2)样本数据的采集过程中变量观测值的观测误差。

(3)理论模型设定的误差。

(4)其他随机因素。

模型式(1.6)清楚地表达了变量 $x_1, x_2, \cdots, x_p$ 与随机变量 y 的相关关系，它由两部分组成：一部分是确定性函数关系，由回归函数 $f(x_1, x_2, \cdots, x_p)$ 给出；另一部分是随机误差项 ε。由此可见模型式(1.6)准确地表达了相关关系既有联系又不确定的特点。

当模型式(1.6)中回归函数为线性函数时，即有

$$y = \beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon \tag{1.7}$$

式中，$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 为未知参数，常称为回归系数。线性回归模型的“线性”是针对未知参数 $\beta_i (i = 0, 1, 2, \cdots, p)$ 而言的。回归解释变量的线性是非本质的，因为解释变量是非线性时，常可以通过变量的替换把它转化成线性的。

如果 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)$ $(i = 1, 2, \cdots, n)$ 是式(1.7)中变量 $(x_1, x_2, \cdots, x_p; y)$ 的一组观测值，则线性回归模型可表示为

$$y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}+\varepsilon_i, \quad i = 1, 2, \cdots, n \tag{1.8}$$

为了估计模型参数的需要，古典线性回归模型通常应满足以下几个基本假设。

(1)解释变量 $x_1, x_2, \cdots, x_p$ 是非随机变量，观测值 $x_{i1}, x_{i2}, \cdots, x_{ip}$ 是常数。

(2)等方差及不相关的假定条件为

$$\begin{cases} E(\varepsilon_i)=0, & i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2, & i=j \\ 0, & i\neq j\end{cases} & i,j=1,2,\cdots,n \end{cases}$$

这个条件称为高斯–马尔柯夫(Gauss-Markov)条件，简称 G-M 条件。在此条件下，便可以得到关于回归系数的最小二乘估计及误差项方差 σ^2 估计的一些重要性质，如回归系数的最小二乘估计是回归系数的最小方差线性无偏估计等。

(3)正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), & i=1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{相互独立} \end{cases}$$